Predicting stroke

"Saba_Alemayehu"

6/9/2022

```
strokeDfv6<- read.csv("healthcare-dataset-stroke-data.csv")
class(strokeDfv6)
```

```
## [1] "data.frame"
```

Exploratory Data Analysis

check name of the columns

```
colnames(strokeDfv6)
```

```
##  [1] "id"                "gender"            "age"
##  [4] "hypertension"      "heart_disease"     "ever_married"
##  [7] "work_type"         "Residence_type"    "avg_glucose_level"
## [10] "bmi"               "smoking_status"    "stroke"
```

```
#str(strokeDfv6)
```

5110 obs. of 12 variables

```
#summary(strokeDfv6)
```

```
head(strokeDfv6)
```

```
##      id gender age hypertension heart_disease ever_married     work_type
## 1  9046   Male  67            0             1          Yes       Private
## 2 51676 Female  61            0             0          Yes Self-employed
## 3 31112   Male  80            0             1          Yes       Private
## 4 60182 Female  49            0             0          Yes       Private
## 5  1665 Female  79            1             0          Yes Self-employed
## 6 56669   Male  81            0             0          Yes       Private
##   Residence_type avg_glucose_level  bmi  smoking_status stroke
## 1          Urban            228.69 36.6 formerly smoked      1
## 2          Rural            202.21  N/A    never smoked      1
## 3          Rural            105.92 32.5    never smoked      1
## 4          Urban            171.23 34.4          smokes      1
## 5          Rural            174.12   24    never smoked      1
## 6          Urban            186.21   29 formerly smoked      1
```

```
#tail(strokeDfv6)
```

```
#apply(strokeDfv6,2,class)
```

encoding the variables into numeric and factors

```
strokeDfv6$id<-as.numeric(strokeDfv6$id)
strokeDfv6$age<-as.numeric(strokeDfv6$age)
strokeDfv6$bmi <- as.numeric(strokeDfv6$bmi)
```

```
## Warning: NAs introduced by coercion
```

```
strokeDfv6$avg_glucose_level<-as.numeric(strokeDfv6$avg_glucose_level)


strokeDfv6$hypertension<-as.numeric(as.factor(strokeDfv6$hypertension))
strokeDfv6$heart_disease<-as.numeric(as.factor(strokeDfv6$heart_disease))

strokeDfv6$gender <- as.factor(strokeDfv6$gender)
strokeDfv6$work_type<-as.factor(strokeDfv6$work_type)
strokeDfv6$ever_married <- as.factor(strokeDfv6$ever_married)
strokeDfv6$Residence_type <- as.factor(strokeDfv6$Residence_type)
strokeDfv6$smoking_status <- as.factor(strokeDfv6$smoking_status)
strokeDfv6$stroke <- as.factor(strokeDfv6$stroke)
```

check the missing values

```
apply(strokeDfv6,2,function(x)sum(is.na(x)))
```

```
##                 id             gender                age       hypertension
##                  0                  0                  0                  0
##      heart_disease       ever_married          work_type     Residence_type
##                  0                  0                  0                  0
## avg_glucose_level                bmi     smoking_status             stroke
##                  0                201                  0                  0
```

check the frequency of each factor variables

```
table(strokeDfv6$gender)
```

```
##
## Female   Male  Other
##   2994   2115      1
```

```
table(strokeDfv6$hypertension)
```

```
##
##    1    2
## 4612  498
```

```
table(strokeDfv6$heart_disease)
```

```
##
##    1    2
## 4834  276
```

```
table(strokeDfv6$ever_married)
```

```
##
##   No  Yes
## 1757 3353
```

```
table(strokeDfv6$work_type)
```

```
##
##      children      Govt_job  Never_worked       Private Self-employed
##           687           657            22          2925           819
```

```
table(strokeDfv6$Residence_type)
```

```
##
## Rural Urban
##  2514  2596
```

```
table(strokeDfv6$smoking_status)
```

```
##
## formerly smoked    never smoked          smokes         Unknown
##             885            1892             789            1544
```

```
table(strokeDfv6$stroke)
```

```
##
##    0    1
## 4861  249
```

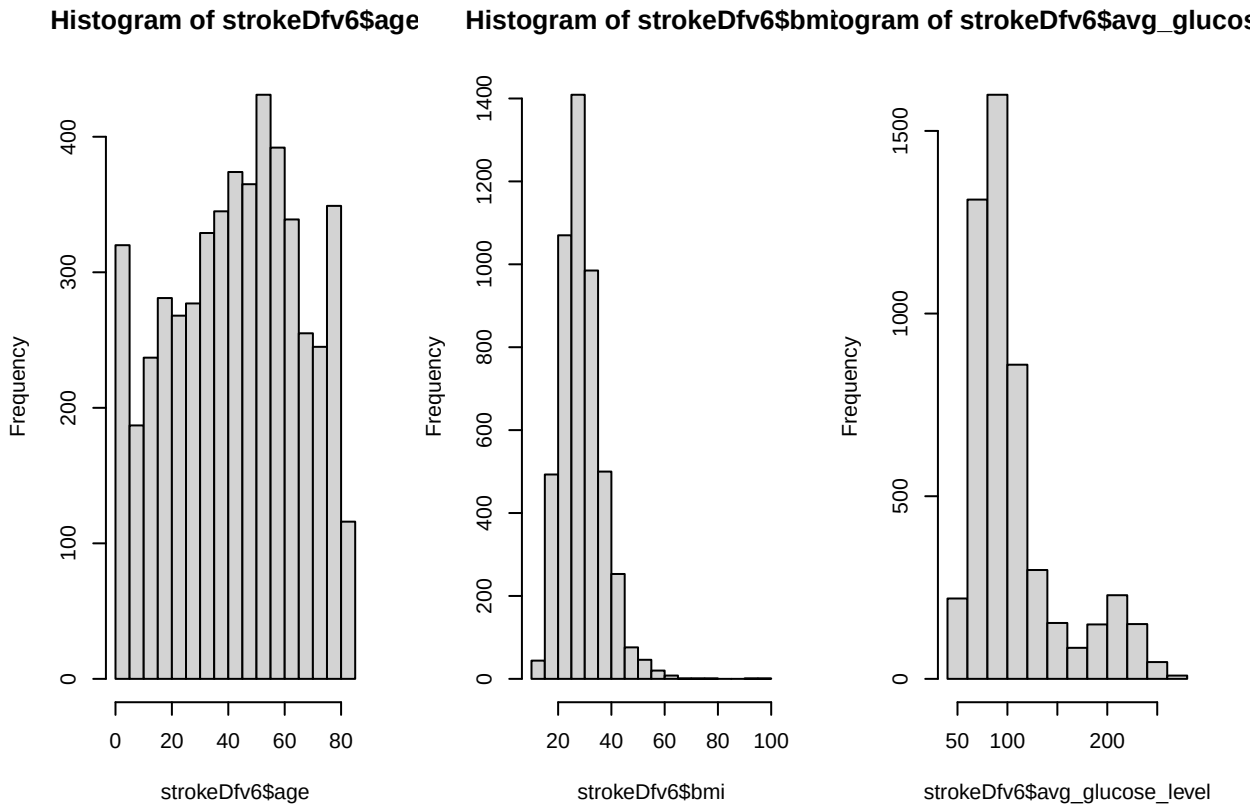
Data Visualization

Variables need to be converted to dummy variables gender, ever_married, work_type, Residence_type, smoking_status

Variables need to be converted to numeric id, age, hypertension, heart_disease, avg_glucose_level, bmi

response variable is stroke stroke should be converted into factor

```
# histogram of numeric variables
par(mfrow=c(1,3))
hist( strokeDfv6$age )
hist( strokeDfv6$bmi )
hist( strokeDfv6$avg_glucose_level )
```




```
# Visualizing categorical predictors
```

convert factors into dummy variables

```
#create dummy variable
dummy_var<-dummyVars(~gender+ever_married + Residence_type+smoking_status, data=strokeDfv6,fullRank = T)

dummy_df<-data.frame(predict(dummy_var,strokeDfv6))
#head(dummy_df)

# combine origional data and dummy data frame

combinedDf<-cbind(dummy_df,strokeDfv6)
head(combinedDf)
```

```
##   gender.Male gender.Other ever_married.Yes Residence_type.Urban
```

```
## 1           1            0               1                    1
## 2           0            0               1                    0
## 3           1            0               1                    0
## 4           0            0               1                    1
## 5           0            0               1                    0
## 6           1            0               1                    1
##   smoking_status.never.smoked smoking_status.smokes smoking_status.Unknown
## 1                           0                    0                     0
## 2                           1                    0                     0
## 3                           1                    0                     0
## 4                           0                    1                     0
## 5                           1                    0                     0
## 6                           0                    0                     0
##      id gender age hypertension heart_disease ever_married     work_type
## 1  9046   Male  67            1             2          Yes       Private
## 2 51676 Female  61            1             1          Yes Self-employed
## 3 31112   Male  80            1             2          Yes       Private
## 4 60182 Female  49            1             1          Yes       Private
## 5  1665 Female  79            2             1          Yes Self-employed
## 6 56669   Male  81            1             1          Yes       Private
##   Residence_type avg_glucose_level  bmi  smoking_status stroke
## 1          Urban            228.69 36.6 formerly smoked      1
## 2          Rural            202.21   NA    never smoked      1
## 3          Rural            105.92 32.5    never smoked      1
## 4          Urban            171.23 34.4          smokes      1
## 5          Rural            174.12 24.0    never smoked      1
## 6          Urban            186.21 29.0 formerly smoked      1
```

```
# remove the redundunt column

stroke_dm <-dplyr::select(combinedDf, -c('gender',
                                         'ever_married',
                                         'Residence_type',
                                         'smoking_status'))

head(stroke_dm)
```

```
##   gender.Male gender.Other ever_married.Yes Residence_type.Urban
## 1           1            0                1                    1
## 2           0            0                1                    0
## 3           1            0                1                    0
## 4           0            0                1                    1
## 5           0            0                1                    0
## 6           1            0                1                    1
##   smoking_status.never.smoked smoking_status.smokes smoking_status.Unknown
## 1                           0                    0                     0
## 2                           1                    0                     0
## 3                           1                    0                     0
## 4                           0                    1                     0
## 5                           1                    0                     0
## 6                           0                    0                     0
##      id age hypertension heart_disease     work_type avg_glucose_level  bmi
## 1  9046  67            1             2       Private            228.69 36.6
## 2 51676  61            1             1 Self-employed            202.21   NA
```

```
## 3 31112  80            1             2       Private            105.92 32.5
## 4 60182  49            1             1       Private            171.23 34.4
## 5  1665  79            2             1 Self-employed            174.12 24.0
## 6 56669  81            1             1       Private            186.21 29.0
##   stroke
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

Data Spliting and data preprocessing

```
#Data partitioning
set.seed(100)
trainingRows<-createDataPartition(stroke_dm$stroke, p=0.8,list=FALSE)
strokeXtrain<-stroke_dm[trainingRows,]
strokeXtest<-stroke_dm[-trainingRows,]

strokeYtrain<-stroke_dm$stroke[trainingRows]
strokeYtest<-stroke_dm$stroke[-trainingRows]

#str(strokeXtrain)
#str(strokeXtest)
```

pre process the training data

#Remove the columns if not used

```
# id, and work type
```

Check for zero variance predictors

```
#check zero variance variables for train data set

stZero_col<- nearZeroVar(strokeXtrain)
#str(stZero_col)
strokeXtrainNZ<-strokeXtrain[,-stZero_col]
strokeXtestNZ<-strokeXtest[,-stZero_col]
#str(strokeXtrainNZ)
#str(strokeXtestNZ)
```

Impute the missing value

```
trainimp<-preProcess(strokeXtrainNZ,"knnImpute")
strokeTrainpr<-predict(trainimp,strokeXtrainNZ)
strokeTestpr<-predict(trainimp,strokeXtestNZ)
```

check for high correlation

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
#corrplot::corrplot(cor(strokeTrainpr))
```

Develop a model